Activation Functions in Neural Networks

Introduction

Activation functions are crucial components of neural networks, introducing non-linearity into the model. They determine the output of a neuron and significantly influence the training dynamics, expressiveness, and performance of the network. This document explores the most commonly used activation functions, detailing their mathematical formulations, advantages, disadvantages, and typical use cases.

Common Activation Functions

1. Sigmoid Function

Definition: The sigmoid function is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Graph:

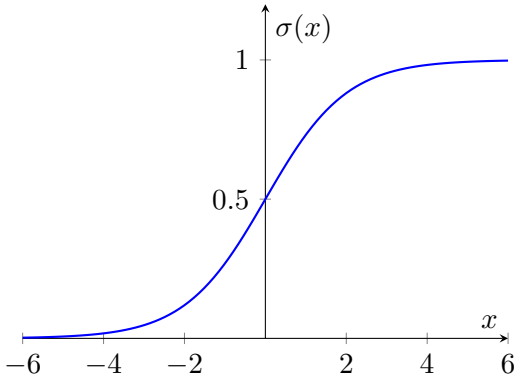

Figure 1: Sigmoid Activation Function

Pros:

- Smooth and differentiable.

- Outputs values in the range (0, 1), making it suitable for probability modeling.

Cons:

- Suffers from the vanishing gradient problem.
- Non-zero centered outputs can slow down convergence.

2. Hyperbolic Tangent (Tanh) Function

Definition: The tanh function is defined as:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Graph:

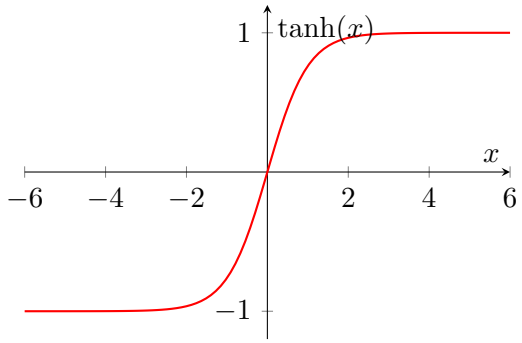

Figure 2: Tanh Activation Function

Pros:

- Outputs values in the range (-1, 1), which is zero-centered.

Cons:

- Suffers from the vanishing gradient problem.

3. Rectified Linear Unit (ReLU)

Definition: The ReLU function is defined as:

$$\text{ReLU}(x) = \max(0, x)$$

Graph:

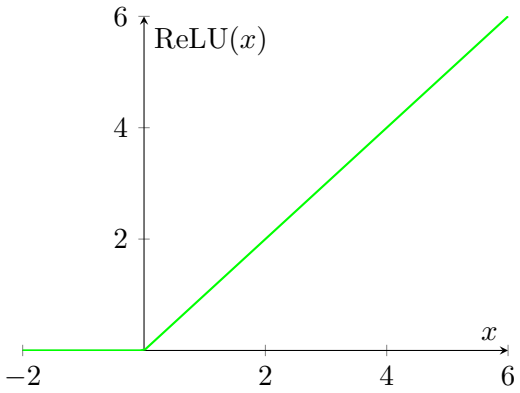

Figure 3: ReLU Activation Function

Pros:

- Efficient and avoids vanishing gradients.

Cons:

- Can suffer from the dying ReLU problem.

4. Leaky ReLU

Definition: The Leaky ReLU function introduces a small slope for negative inputs:

$$\text{Leaky ReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{if } x < 0 \end{cases}$$

where α is a small constant, typically 0.01.

Graph:

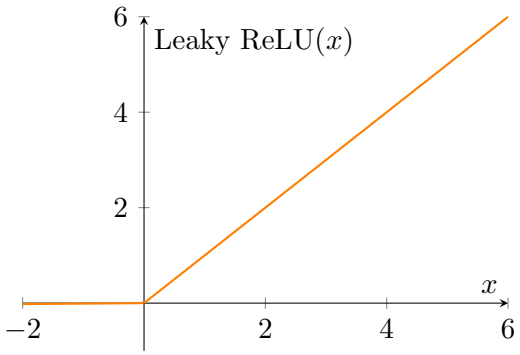

Figure 4: Leaky ReLU Activation Function

Pros:

- Mitigates the dying ReLU problem.

5. Softmax

Definition: The softmax function converts a vector of raw scores into probabilities:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

Pros:

- Outputs a probability distribution.

6. Gaussian Error Linear Unit (GELU)

Definition: The GELU function is approximated as:

$$\text{GELU}(x) \approx x \cdot \sigma(1.702x)$$

Graph:

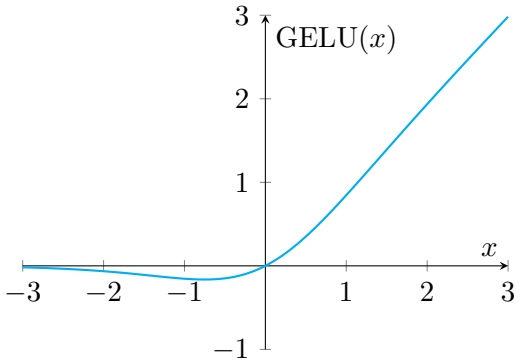

Figure 5: GELU Activation Function (Approximation)

7. Swish

Definition: The Swish function is defined as:

$$\text{Swish}(x) = \frac{x}{1 + e^{-x}}$$

Graph:

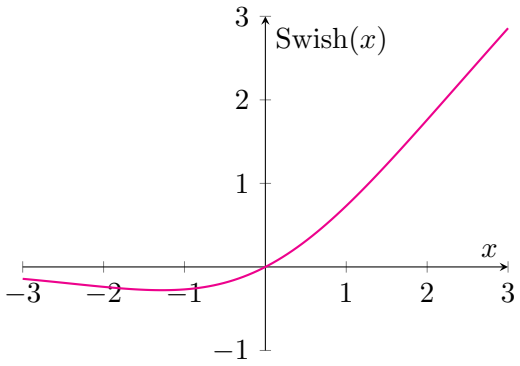

Figure 6: Swish Activation Function

Conclusion

The choice of activation function is critical for neural networks. Each activation function is suited for specific tasks, and understanding their behavior can improve model performance.